Article:
Sharma, Rajan, Gu, Yuanyuan, Sinha, Kompal et al. (2 more authors) (2019) Mapping the Strengths and Difficulties Questionnaire onto the Child Health Utility 9D in a large study of children. Quality of life research. 2429–2441. ISSN 1573-2649

Mapping the Strength and Difficulties Questionnaire onto the Child Health Utility 9D in a large study of children

Running heading: Mapping the SDQ onto the CHU9D

Rajan Sharma*
Ph.D. candidate
ORCID ID: 0000-0002-4474-5296
Macquarie University Centre for the Health Economy
Faculty of Business and Economics
Level 1, 3 Innovation Road, Macquarie University, NSW, 2109, Australia
Email: rajan.sharma1@students.mq.edu.au
Phone: +61 2 9850 4768

Yuanyuan Gu, Ph.D.
Marie Skłodowska-Curie Fellow
ORCID ID: 0000-0002-3816-9106
Department of Economics and Related Studies
University of York, Heslington Road, York, YO10 5DD, United Kingdom
Senior Research Fellow
Macquarie University Centre for the Health Economy
Faculty of Business and Economics
Level 1, 3 Innovation Road, Macquarie University, NSW, 2109, Australia

Kompal Sinha, Ph.D.
Senior Lecturer
ORCID ID: 0000-0003-4318-6100
Department of Economics
Macquarie University, NSW, 2109, Australia

Mona Aghdaee, Msc
Research Fellow
ORCID ID: 0000-0002-2570-1685
Macquarie University Centre for the Health Economy
Faculty of Business and Economics
Level 1, 3 Innovation Road, Macquarie University, NSW, 2109, Australia

Bonny Parkinson, Ph.D.

Senior Research Fellow

ORCID ID: 0000-0002-2137-0962

Macquarie University Centre for the Health Economy

Faculty of Business and Economics

Level 1, 3 Innovation Road, Macquarie University, NSW, 2109, Australia

*Corresponding author

Abstract

Purpose

Non-preference-based measures cannot be used to directly obtain utilities but can be converted to preference-based measures through mapping. The only mapping algorithm for estimating Child Health Utility-9D (CHU9D) utilities from Strengths and Difficulties Questionnaire (SDQ) responses has limitations. This study aimed to develop a more accurate algorithm.

Methods

We used a large sample of children (n= 6,898), with negligible missing data, from the Longitudinal Study of Australian Children. Exploratory factor analysis (EFA) and Spearman's rank correlation coefficients were used to assess conceptual overlap between SDQ and CHU9D. Direct mapping (involving seven regression methods) and response mapping (involving one regression method) approaches were considered. The final model was selected by ranking the performance of each method by averaging the following across ten-fold cross-validation iterations: mean absolute error (MAE), mean squared error (MSE), and MAE and MSE for two subsamples where predicted utility values were <0.50 (poor health) or >0.90 (healthy). External validation was conducted using data from the Child and Adolescent Mental Health Services study.

Results

SDQ and CHU9D were moderately correlated (ρ=-0.52, $p<0.001$). EFA demonstrated that all CHU9D domains were associated with four SDQ subscales. The best-performing model was the Generalised Linear Model with SDQ items and gender as predictors (full sample MAE: 0.1149; MSE: 0.0227). The new algorithm performed well in the external validation.

Conclusions

The proposed mapping algorithm can produce robust estimates of CHU9D utilities from SDQ data for economic evaluations. Further research is warranted to assess the applicability of the algorithm among children with severe health problems.

Keywords: CHU9D; SDQ; mapping; utility

Introduction

Cost-utility analysis (CUA) is increasingly being used to make decisions regarding resource allocation[1]. In CUAs health outcomes are usually measured in terms of gain in Quality-adjusted life-years (QALYs)[1], although they can also be measured in terms of Disability Adjusted Life Years (DALYs) and Health-Adjusted Life Years (HALYs). Economic evaluation guidelines from regulatory and reimbursement agencies recommend the estimation of QALYs when evaluating healthcare technologies and services[2-4]. Using QALY as the health outcome measure is preferred for a number of reasons. First, QALYs combine both duration in a health state and quality of life, measured using preference weights (or utilities) associated with that health state, in a single metric[5,6]. Second, availability of willingness-to-pay thresholds, such as incremental cost per QALY gained, makes it possible for the decision makers to decide whether an intervention should be funded[7]. Last, it is possible to compare QALYs across different treatments and diseases and rank them in terms of their effectiveness and cost-effectiveness[8]. Consequently, it is easier for policy-makers to make funding decisions.

Estimation of QALYs is based on health state utility values[1] that reflect the preferences or desirability expressed by an individual for a particular health state[10,11]. These utilities are obtained by administering a multi-attribute utility instrument[2] (MAUI), such as the EuroQol five-dimensions (EQ-5D)[12,13], Short-Form 6D (SF-6D)[14,15] or other generic preference-based measures (PBMs) to patients to describe their health state. Health states are then valued using pre-existing scoring algorithm obtained from techniques such as time trade-off[16] and standard gamble[17].

MAUIs may not be sensitive enough to distinguish between clinical severity levels[18,19]. They are not always included in surveys or clinical studies. Instead, non-PBMs are collected[1]. Non-PBMs do not have a preference-based algorithm and thus cannot be used to estimate utilities. However, they can be mapped[3] to PBMs, thus allowing the estimation of utility values[20]. Guidelines from regulatory and reimbursement agencies recognise mapping as a technique to estimate utilities[21-23]. A review by Kearns et al. found that mapping algorithms were used in around a quarter of the health technology appraisals for the National Institute for Health and Care Excellence (NICE) in the United Kingdom[24].

The present study is concerned with predicting the Child Health Utility 9D (CHU9D) utility values from the Strengths and Difficulties Questionnaire (SDQ). Only one such mapping algorithm exists. It was developed by Furber et al. using an Australian sample[25]. However, the published version of the algorithm was developed using an older version of the CHU9D instrument[26] and thus can no longer be used.[4] Nevertheless, the sample was small (N=200) and from a narrow population - children receiving mental health services. It also only considered the ordinary least squares (OLS) method when developing the mapping algorithm, which may not accommodate the potential features of the utility scores' distribution such as multimodality, skewness, and a high proportion equalling one[27]. Finally, it used proxy-reported data based on caregivers' responses to the SDQ and CHU9D.

This study aimed to develop a new and more accurate mapping algorithm for estimating CHU9D utilities from SDQ, using a larger survey of Australian children and applying a variety of mapping approaches. It also used self-reported SDQ and CHU9D data, which may measure Health-related Quality of Life (HRQoL) better than proxy-reported data[28].

---

[1] Utilities are measured on a cardinal scale, anchored at 0 (death) and 1 (full health)[5,9,1]. Utilities less than zero are possible if a health state is considered to be worse than death[1].

[2] MAUIs are a health-related quality of life questionnaires which are associated with an algorithm to convert responses to the included questions into utilities.

[3] Mapping is the process of establishing a statistical relationship between a non-PBM and a PBM using regression techniques.

[4] It can certainly be updated using the new version of the CHU9D instrument but to our best knowledge such an updated algorithm is not publicly available. In this study we obtained a subsample of the original data used in Furber et al. from the lead author and updated their algorithm to be used as a comparator of our new algorithm in the external validation.

Methodology

Data

This study used the Longitudinal Study of Australian Children (LSAC) dataset. This is a nationally representative survey of the Australian children involving biennial data collection, through 2004 (Wave 1) to 2014 (Wave 6)[29]. Wave 1 of the LSAC was initiated by collecting data for a birth (B) cohort (i.e. <1 year of age, n= 5,107) and a kindergarten (K) cohort (i.e. 4-5 years of age, n= 4,983). In Wave 6 in 2014 CHU9D and SDQ questionnaires were administered to the children. Thus, the Wave 6 of the LSAC from both B (10-11 years old) and K (14-15 years old) cohorts was used to map SDQ onto CHU9D. In total, the sample size was 7,301 (B-cohort, N= 3,764 and K-cohort, N= 3,537).

The amount of missing data was small relative to the large sample size, and thus unlikely to cause significant bias[30,31]. There were 370 (0.06%) missing values for the SDQ and 400 (0.05%) missing values for the CHU9D in the whole sample (N= 7,301). Thus, no statistical measures, such as multiple imputations, were applied to address the missing data problem. The final sample consisted of 6,898 children.

External validation was conducted using Child and Adolescent Mental Health Service (CAMHS) dataset (N=103)[25]. The full sample of the CAMHS dataset (N=200) was previously used to develop a mapping algorithm between SDQ and CHU9D by Furber et al. (2014). Unfortunately, only a subset of the data could be obtained for our external validation.

Estimation and validation samples

A ten-fold cross-validation approach was used to assess the predictive performance of the models[32]. This involved dividing the original sample from the LSAC dataset into ten equal-sized sub-samples using a random number generator. In each iteration, nine of the ten groups (90% of the dataset) were allocated to the 'estimation sample' and the remaining group (10% of the dataset) was used as the 'validation sample'. This process was repeated ten times in order to ensure that each of the ten subgroups was used in both estimation and validation iterations.

Source and target measures

Source measure: Strengths and difficulties questionnaire (SDQ)
The SDQ was designed to describe HRQoL (over the past six months) of children and adolescents aged 2 through 17 years old[30]. It includes 25 items within five subscales (emotional symptoms, conduct problems, hyperactivity, peer problems, and pro-social). Each item allows a response from 1 to 3 representing the presence or lack of problems in terms of "not true", "somewhat true" or "certainly true". Individual subscales range between 0 and 10, with higher scores indicating poorer functioning or poorer HRQoL. The total score is obtained by adding scores from four subscales as per SDQ guidelines[33]. The total SDQ score ranges from 0 to 40 with a higher score associated with poorer HRQoL. Descriptive information about each SDQ subscale and item is presented in the electronic supplementary materials (ESM_1, Table 1).

Target measure: Child Health Utility 9D (CHU9D)
The CHU9D is a MAUI specifically designed to estimate utilities experienced by children and adolescents[34]. It includes nine domains (worried, sad, pain, tired, annoyed, schoolwork/homework, sleep, daily routine, and activities). Each domain allows a response from 1 (no problems) to 5 (severe problems) and is assessed as of 'today'. In comparison to other MAUIs that can be used to measure utilities in young populations, the CHU9D does not involve the adaptation of an existing adult instrument because it was developed with the young population since its inception[35]. Descriptive information about each CHU9D domain is presented in ESM_1, Table 2.

Utilities were estimated by applying an Australian adolescent-specific (11-17 years) scoring algorithm to the CHU9D responses in both LSAC and external validation datasets[26]. These utilities can range between -0.1059 to 1. The target measure (CHU9D utilities or individual domains) was predicted using different source measures (SDQ total scores, domains or items).

Statistical analysis

All statistical analyses were carried out in STATA 15.1[36]. This study follows the Mapping onto Preference-Based Measures Reporting Standards (MAPS) checklist[37] (see ESM_1, Table 3).

Exploratory data analysis

The precision of the mapping approaches relies on the extent of overlapping between the source and target measures[6,38,39]. The correlation between total SDQ scores and the CHU9D utilities was evaluated using Spearman's correlation coefficient, with an associated 95% confidence interval (CI) computed using 1,000 bootstrap iterations. The visual relationship between the measures was explored using a scatterplot. Additionally, Kernel density plots were drawn to visually assess the distribution of those measures.

The inclusion of square or cube forms of predictors were considered to account for any non-linear relationships with the CHU9D utilities. They were however not included in the final models due to multicollinearity. Variance inflation factor in excess of 10 was considered as an indication of multicollinearity[40].

Exploratory factor analysis (EFA) was conducted to understand if the SDQ subscales and CHU9D domains could be described by the same latent constructs or factors. Kaiser-Meyer-Olkin (KMO) test measure was used to determine sampling adequacy for EFA[41]. Bartlett test for sphericity was used to test the null hypothesis that variances between variables were equal[42]. Variances of the factors and correlations of a domain or a subscale with a factor were examined using "eigenvalues" and "factor loadings". Factors were oblique-rotated to allow for possible correlation between domains or subscales. Factors were retained when eigenvalues exceeded one[43]. Factor loadings exceeding 0.3 were considered "meaningful" in suggesting that SDQ subscales and CHU9D domains were capturing the same underlying construct[44].

Modelling approaches – direct mapping

We adopted direct and indirect approaches to mapping. The direct mapping approach involved directly regressing CHU9D utilities on the SDQ total, subscales, or item scores. Predicted utilities greater than 1 were rounded to 1.

Four sets of predictor were considered:

Predictor set 1
Total SDQ score and Gender

Predictor set 2
SDQ subscales score and Gender

Predictor set 3
SDQ items (categorical) and Gender

Predictor set 4
SDQ items (continuous) and Gender

SDQ subscales include emotional symptoms, conduct problems, hyperactivity, peer problems, and pro-social behaviours; SDQ subscales score except 'pro-social' were added to derive *Total SDQ score*; and *SDQ item* includes 25 SDQ items (detailed in ESM_1, Table 1). Gender (Female= 1) was also included in all models. The exclusion of the 'pro-social' subscale was initially considered because it was very weakly correlated with utilities. However, including this variable in the OLS model slightly improved the adjusted R^2 (from 0.3458 to 0.3493) and overall mean squared error (MSE) of the model (from 0.1545 to 0.1540). It was thus included as one of the predictors in Predictor set 2.

In Predictor set 1 the SDQ total scores were included as continuous independent variables. In Predictor set 2 all SDQ subscales scores were included as continuous independent variables. In Predictor set 3 all SDQ items were included as categorical independent variables. In Predictor set 4 SDQ items were included as continuous independent variables. Among individual characteristics, only gender was included as an independent variable to ensure the generalisability of the mapping algorithm. Age was also considered but not included in the final predictor sets as it was not statistically significant.

We considered several regression methods as there was no consensus in the literature regarding which regression method would best accommodate the unusual distribution of the utility values. The methods included: ordinary least squares (OLS), generalised linear model (GLM), extended estimating equations (EEE), zero-one-

inflated beta regression (ZOIB), Tobit model, censored least absolute deviation (CLAD) and finite mixture model (FMM).

OLS[45] was used as it is the most commonly used method in mapping literature[6]. GLM[46,47] was used to predict disutilities (i.e.,1-utilities)[5] as it has two distinct advantages over the OLS: (i) it accommodates the nonlinear relationship between predictors and the dependent variable (through the link function) and potential heteroscedasticity (through the variance function or family distribution); and (ii) it provides consistent estimates even if the variance function is incorrectly specified (given the link function is correctly specified). The modified Parks test (MPT)[48] was used to identify the family distribution based on lowest χ^2 value. The Pregibon Link Test[49], the Pearson Correlation test[50] and the modified Hosmer-Lemeshow test[51] were used to determine an appropriate link function. A link function is deemed to fit well when all three tests yield insignificant p-values[52]. We found the family distribution as Poisson and the link function as a power function with power at 0.75) for predictor sets 1, 2 and 4 and power at 0.50 for predictor set 3 to be appropriate for the GLMs. EEE is a flexible version of GLM. It directly estimates the family distribution (represented by the value of the θ_2 parameter) and link function (represented by the value of the λ parameter)[53]. The θ_2 and λ values in the EEE models were found to be consistent with the family and link function parameter values determined in the GLM. Thus the use of EEE was helpful in confirming the appropriate family and link functions. The ZOIB[54] was used as it was deemed appropriate for modelling proportional dependent variable containing 0's and/or 1's. The Tobit model [55-57] was used to predict utilities in the presence of censoring. The CLAD[58], a median based model, was used to deal with observed utilities with ceiling effects, heteroscedasticity, and skewness[59]. The FMM[60] was used to deal with the multimodal and skewed distribution of CHU9D utilities[61,62]. FMM was expected to capture the unobserved heterogeneity of effects for individuals who belong to different classes or components[63,60].

Modelling approaches – indirect mapping

An indirect or 'response' mapping approach estimates the predicted probabilities for each level of the CHU9D domains and converts them into utilities using the corresponding Australian algorithm[64] through the 'Expected value approach' technique[65]. Multinomial logistic regression (MLOGIT)[66] was applied over Ordinal logit regression (OLOGIT) as the Brant test[6] found that the 'parallel odds assumption' (one of the main assumptions of the OLOGIT) did not hold true for all CHU9D domains. Gender (Female =1) was also included. The same set of predictors used in the direct mapping approach were considered.

Four predicted probabilities were estimated, as each CHU9D domain has five levels, using Equation 1:

$$Pr(CHU9D\ domain_i = m|Z_i = z) = \frac{exp(z_{mi})}{1 + \sum_{m=2}^{M} exp\,(z_{mi})} \qquad \text{Equation 1}$$

where m = 2, 3, ……, M for each level of CHU9D domains (worried, sad, pain, tired, annoyed, schoolwork/homework, sleep, daily routine, and activities), and Z represents covariates.

Probabilities for the reference category were estimated using Equation 2:

$$Pr(CHU9D\ domain_i = 1|Z_i = z) = \frac{1}{1 + \sum_{m=2}^{M} exp\,(z_{mi})} \qquad \text{Equation 2}$$

Measures of predictive accuracy

The application of each regression method was combined with each predictor set, which resulted in 32 candidate models. Predictive accuracy of the models was compared by averaging the following measures across ten-fold cross-validation iterations: i) Mean Absolute Error (MAE)[7], ii) MSE[8], iii) MAE/MSE using a subsample where observed utility scores were greater than 0.90 (representing the healthy group of children), and iv) MAE/MSE using a subsample where observed utility scores were less than 0.50 (representing the group of children with poor HRQoL). Ten-fold cross-validation method was used to assess goodness-of-fit and to avoid the risk of

[5] For GLM and EEE the disutility was used as the outcome variable to avoid non-negative values.

[6] Brant test was used to check proportional odds assumption to determine the ordered nature of CHU9D domains. Parallel assumption or proportional odds assumption is an assumption of an ordered logit model. This assumption assumes that coefficients between different categories of the dependent variable are equal.

[7] The MAE was calculated as mean of the absolute values of the difference between the observed and predicted CHU9D utilities.

[8] The MSE were computed as the mean squared differences between the predicted and observed CHU9D utilities.

over-fitting[67]. MAEs and MSEs are commonly used as measures of predictive accuracy. Two subsamples were chosen because predicted utilities are prone to underestimating the lower utilities (around observed utility <0.5) and overestimating in the upper extreme (around observed utility >0.9).

Models were ranked according to each of the above criteria resulting in a number of rankings. These were then averaged to produce the overall ranking. The lower the overall ranking, the better the performance of the model.

A similar ranking procedure was adopted during sensitivity analyses. Sensitivity analyses were conducted to test the robustness surrounding the choice of the best performing model by ranking models based on average results from cross-validation using: i) only MAE; ii) only MSE; and iii) MAE/MSE across different ranges of observed utilities (that is, MAE/MSE for utility range <0.2, 0.2 to 0.4, 0.4-0.6, 0.6-0.8, and 0.8-1).

Final algorithm

The final mapping algorithm was suggested based on the best performing model using the entire sample. STATA and Excel tools were produced to enable future mapping exercise from the SDQ to the CHU9D. Robust standard errors were also reported.

External validation

One of the key differences between the sample used in this study and Furber et al. is that the latter (a) was based on patients receiving mental health services, namely CAMHS and; (b) it used an older version of the CHU9D tariffs to calculate utilities[26], while this study utilised the latest version of the CHU9D tariffs[64]. We conducted external validation of our proposed mapping algorithm using a subsample of the CAMHS dataset provided by the lead author of Furber et al. First, we applied the latest version of CHU9D tariffs to obtain observed CHU9D utilities in the CAMHS dataset. Then, we estimated an OLS regression with SDQ subscales as predictors and obtained coefficients to be used as new or modified mapping algorithm for Furber et al. Finally, we used our proposed algorithm in the CAMHS dataset and compared the findings in terms of mean, MAEs, MSEs, and their 95% CIs.

Results

Descriptive statistics

Table 1 presents the socio-demographic characteristics of the LSAC and external validation samples. Children in LSAC were comparatively healthier than those in CAMHS [CHU9D utilities over 0.90: 36.9% versus 13.6%; SDQ total score: 9.94 versus 20.31; subscale scores (hyperactivity: 3.71 versus 4.64; emotional symptom: 2.91 versus 5.38; peer problems: 1.68 versus 5.84; and conduct problems: 1.63 versus 4.45)]. Samples were similar in terms of mean age and gender distribution.

<TABLE 1 HERE>

Fig.1 illustrates the Kernel density plot of the observed CHU9D utilities and the SDQ total scores. These plots along with the Shapiro Wilks test for normality rejected the null hypothesis of normally distributed CHU9D utilities ($p<0.001$).

<FIGURE 1 HERE>

A moderately strong statistically significant correlation between the observed utilities and the total SDQ was observed (Spearman's rho (ρ) = -0.52; 95% CI: -0.54 to -0.50; $p<0.001$). Such a correlation was also observed in the scatterplot of utilities and total SDQ scores. The highest correlation existed between the utilities and SDQ subscale “emotional symptoms” (ρ=-0.51, $p<0.001$) (see ESM_1, Fig. 1 and Table 4).

The sample adequacy test prior to EFA found the sample to be appropriate for factor analysis (KMO: 0.88) [41,68]. The Bartlett test for sphericity rejected the null hypothesis that variances between variables were equal (χ^2 = 34020.69, $p<0.001$). Consequently, the sample was determined to be fit for factor analysis. EFA resulted in

one key factor with meaningful loadings on all SDQ subscales except pro-social subscale, as well as all nine CHU9D domains. This overlap in the same factor suggests that all nine CHU9D domains possibly capture the similar latent construct as the four SDQ subscales. These results provided evidence that there is adequate conceptual overlap such that the mapping algorithm would be valid. Results of the EFA are provided in ESM_1, Table 5, Table 6 and Fig. 2.

Predictive accuracy results based on cross-validation

Table 2 presents predictive accuracy results. The direct approach using GLM 3 (i.e. GLM method using predictor set 3) was considered the best regression method based on the overall ranking. The rank of the best performing model did not change during the sensitivity analysis demonstrating the robustness of the results. Results from sensitivity analyses are presented in ESM_1, Table 7.

<TABLE 2 HERE>

MAEs ranged from 0.1151 (MLOGIT 3) to 0.1222 (ZOIB 1). MSEs ranged from 0.0230 (GLM 3 and EEE 3) to 0.0256 (MLOGIT 1). This is equivalent to percentage errors of up to 10%[9] and 2%[10], respectively, of the overall CHU9D range. The difference between the mean of the predicted utilities and the mean of the observed utilities was zero to three decimal places for all OLS, GLM, and EEE regression methods using any of the four predictor sets. Thus, on average, they were able to exactly predict the mean CHU9D utilities for the LSAC population.

The distribution of the predicted utilities indicates that all models overpredicted at the lower extreme and underpredicted at the upper extreme of the observed utilities (see ESM_1, Fig. 3). This was anticipated given the unusual distribution of the observed utilities. However, the prediction seemed to improve when categorical items (Predictor set 3) instead of total SDQ or SDQ subscales were used as explanatory variables. The models performed best when the observed utilities ranged between 0.6 to 0.8 as a high number of predicted utilities will overlap with the observed utilities.

Mapping function

GLM 3 was considered the best performing model based on the overall ranking (see ESM_1, Fig. 4). Coefficients from GLM 3 are presented in Table 3. Separate STATA and Excel tools to implement the algorithm are provided in ESM_2 and ESM_3.

<TABLE 3 HERE>

External validation

The mean (95% CIs) observed CHU9D utility value in the CAMHS dataset was 0.6161 (0.5727 to 0.6597). The mean (95% CIs) utility value after applying our algorithm in the CAMHS dataset was 0.6346 (0.6066 to 0.6627). Although the mean observed CHU9D utility value and the one resulting from using our algorithm were not identical, they were similar. Some discrepancy was expected since the CAMHS and LSAC samples are considerably different from each other. The resulting mean (95% CIs) MAE using our algorithm in the CAMHS dataset was lower [0.1522 (95% CI: 0.1294, 0.1750)] than the one using Furber et al.'s procedure in the same dataset [0.1578 (95% CI: 0.1376, 0.1780)]. The corresponding MSEs were similar 0.0366 (95% CI: 0.0275, 0.0457) and 0.0355 (95% CI: 0.0277, 0.0432), respectively.

[9] Formula: $\%\ Error\ for\ GMAE = 100 * \frac{MAE}{(Max(CHU9D\ utilities) - Min(CHU9D\ utilities))}$

[10] Formula: $\%\ Error\ for\ GMSE = 100 * \frac{MSE}{(Max(CHU9D\ utilities) - Min(CHU9D\ utilities))}$

Note: $(Max(CHU9D\ utilities)$=1 and $(Min(CHU9D\ utilities)$= -0.1059

Discussion and conclusion

This study aimed to propose a mapping algorithm to predict CHU9D health state utilities from SDQ. We considered a variety of mapping approaches, each having its own advantages. The ranking system adopted in this study avoids decision-making based on a single criterion. Robustness of ranking was further assessed through sensitivity analyses. The best performing mapping algorithm was able to accurately predict the mean observed utility, with better predictions observed at utilities between 0.6 and 0.8. STATA and Excel tools were produced to enable the use of the algorithm in future economic evaluations.

The predicted utility values after applying the best performing model from this study (GLM 3) to the full sample resulted in the mean utility of 0.7976 (SD= 0.12), ranging from 0.2000 to 0.9600. Using the modified version of the previous algorithm[11] in our study sample resulted in predicted utility values with the mean of 0.8190 (SD= 0.07), ranging from 0.5500 to 0.9500[25]. This suggests that the new algorithm outperformed the previous one in predicting the observed mean (i.e. 0.7976) and range prediction for this dataset. Both algorithms did not perform well in predicting the upper end of the utility distribution but the proposed mapping algorithm is an improvement in this regard. The GLM 3 model also outperformed the modified version of the mapping algorithm from Furber et al. in the prediction of mean CHU9D utilities when respondents were split by their self-reported global health measure (see Fig. 2). A figure showing the predicted utilities from the best model compared to observed utilities across the self-reported global health measure is presented in ESM_1, Fig. 5.

<FIGURE 2 HERE>

The predictive performance of the preferred GLM 3 model (full sample MAE: 0.1146 and MSE: 0.0225) is within the ranges reported by previous mapping studies (MAE: 0.0011 to 0.1900 and MSE: 0.007 to 0.040)[6]. GLMs were also found to outperform other models in other mapping studies[69-71]. GLMs bear many advantages when dealing with highly skewed data and have been proven to be particularly successful in modelling healthcare costs which share similar features with the health utilities[72]. It is therefore somewhat surprising that GLMs have not been more widely considered in mapping studies.

The models that were close to win include Tobit, MLOGIT, and CLAD. However, their high ranks were largely due to their superior performance on the upper tail of the distribution (see Table 2). The censoring mechanism within Tobit and CLAD naturally led to small prediction errors at the top end which was unfortunately offset by their rather poor performance on the lower end. One of the key conditions for using response mapping is that enough responses are needed at all levels in each dimension[20]. MLOGIT's superior performance at the top end and poor performance on the lower end suggested this condition might have not been fulfilled.

These three models were closely followed by OLS which performed relatively well on the lower end but not so well on the upper tail. The models ranked lowest are ZOIB, EEE, and FMM. It is interesting that they are the "flexible" models but the potential over-fitting problem may have caused them to perform well on some regions and not so well on the others. For example, EEE had very good performance based on the full sample MAE or MSE but was among the worst on both ends.

The EEE model has been rarely explored in the mapping literature. This might be due to the fact that the authors of EEE recommended using a fairly large sample size (N=5000)[53]. To our knowledge, only one mapping study used EEE[73]but its sample size was only 772. It used two-fold cross-validation which suggested EEE to be the best performing model for one of its two mapping exercises. Our results did not support this, largely due to that we factored in the performance of the model in different regions of the utility distribution. It is also worth mentioning that EEE does not necessarily outperform a carefully selected GLM whose link and variance functions are chosen via multiple robust tests[72]. But EEE can be estimated as the starting point for selecting the optimal GLM.

The key strengths of this study are as follows. First, this study used a sample from a much larger (n=6,898) and more diverse population. Second, it considered both direct and indirect approaches for the mapping. Within the direct mapping, we looked beyond OLS and compared a range of econometric methods for selecting the best performing model. Using only OLS may not accommodate the potential features of the utility scores' distribution: such as multimodality, skewness, and a high proportion of utilities equalling one[22]. Moreover, unlike many other studies, we also factored in the performance of a model on the upper and lower tails of the

[11]We used new CHU9D tariff to obtain observed CHU9D variable in the CAMHS dataset, ran an OLS regression with SDQ subscales as predictors and obtained coefficients to be used as new or modified mapping algorithm for Furber et al.

utility distribution in order to select one that can also predict well for children with very good and poor HRQOL which is important for economic evaluation. Third, this study used self-reported SDQ and CHU9D data which may measure HRQoL better than proxy-reported data[28]. Fourth, this study performed an external validation of the mapping algorithm by assessing the predictive performance of the algorithm using an external dataset. Our mapping algorithm also captured wider values at extremes (0.3000 to 0.9400 as against 0.3900 to 0.9000 using CAMHS dataset). These results indicate that the proposed algorithm may perform better even among children utilising mental health services compared to Furber et al. We conclude that our mapping algorithm predicted well both in-sample and out-of-sample.

One of the potential limitations of the models was that they did not capture lower utilities. The finding may raise concerns about the applicability of the proposed algorithm to children with severe health conditions. Although this limitation occurs frequently in mapping models[6,74,75], the poor performance in this study was probably due to very few observations at lower extreme. Future studies could attempt to derive more robust estimations for this particular sub-group of children, although it may prove difficult given the small number of children reporting poor health on CHU9D. Researchers willing to adopt this mapping algorithm should carry out sensitivity analyses in economic evaluations surrounding the results obtained by applying the algorithm. Moreover, the recommended algorithm based on the GLM 3 model did not perform well in capturing extremely high utilities (>0.96). Finally, only a relatively small dataset was available for the external validation.

In conclusion, the proposed mapping algorithm is a better predictor of CHU9D utilities. Further research is warranted to assess the performance of the proposed mapping algorithm in a larger sample consisting of children and adolescents with more severe health conditions.

Acknowledgements

Some of the results in this manuscript were presented at 40th Australian Health Economics Society Conference in Hobart, Australia in September, 2018. The authors are grateful to the audience for helpful comments. The authors would like to acknowledge the Department of Social Services and National Centre for Longitudinal Data for reviewing and approving the study protocol and for providing assistance in using the datasets. The authors would like to thank Dr. Gareth Furber for making the CAMHS dataset available for external validation. The authors would also like to acknowledge the feedback from Anirban Basu, Professor of Health Economics at University of Washington.

Authors' contributions

RS, YG, KS, MA and BP contributed to the conception and design of this mapping exercise. RS conducted the statistical analysis. All the contributors contributed to the interpretation of data; drafting the article, revising it critically for the intellectual content and final approval version to be published.

Funding: This paper is a part of a Ph.D. project funded by International Macquarie University Research Excellence Scholarship (iMQRES). Yuanyuan Gu's research is supported by a Marie Sklodowska-Curie Individual Fellowship (No. 740654).

Compliance with Ethical Standards

Conflict of Interest: RS, YG, KS, MA and BP declare that there is no conflict of interest regarding the publication of this article.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors. Department of Social Services and National Centre for Longitudinal Data reviewed and approved the study protocol and made the LSAC datasets available.

References

1. Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L., & Torrance, G. W. (2015). *Methods for the economic evaluation of health care programmes*: Oxford university press.
2. Canadian Agency for Drugs Technologies in Health (2006). Guidelines for economic evaluation of pharmaceuticals: Canada. Ottawa: Canadian Agency for Drugs and Technologies in Health.
3. National Institute for Health and Clinical Excellence (2013). Guide to the methods of technology appraisal 2013.
4. Pharmaceutical Benefits Advisory Committee (2016). Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee (version 5.0). Australian Government Department of Health; 2016.
5. Whitehead, S. J., & Ali, S. (2010). Health outcomes in economic evaluation: the QALY and utilities. *British medical bulletin, 96*(1), 5-21, doi:10.1093/bmb/ldq033.
6. Brazier, J. E., Yang, Y., Tsuchiya, A., & Rowen, D. L. (2010). A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *The European Journal of Health Economics, 11*(2), 215-225.
7. Harris, A. H., Hill, S. R., Chin, G., Li, J. J., & Walkom, E. (2008). The role of value for money in public insurance coverage decisions for drugs in Australia: a retrospective analysis 1994-2004. *Medical Decision Making, 28*(5), 713-722.
8. Neumann, P. J., Cohen, J. T., & Weinstein, M. C. (2014). Updating cost-effectiveness—the curious resilience of the $50,000-per-QALY threshold. *New England Journal of Medicine, 371*(9), 796-797.
9. Shiell, A., Donaldson, C., Mitton, C., & Currie, G. (2002). Health economic evaluation. *Journal of Epidemiology & Community Health, 56*(2), 85-88.
10. Tolley, K. (2009). What are health utilities. *Hayward Medical Communications, London*.
11. Torrance, G. W. (1987). Utility approach to measuring health-related quality of life. *Journal of chronic diseases, 40*(6), 593-600.
12. Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical care*, 1095-1108.
13. Shaw, J. W., Johnson, J. A., & Coons, S. J. (2005). US valuation of the EQ-5D health states: development and testing of the D1 valuation model. *Medical care*, 203-220.
14. Brazier, J., Roberts, J., & Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of health economics, 21*(2), 271-292.
15. Brazier, J. E., & Roberts, J. (2004). The estimation of a preference-based measure of health from the SF-12. *Medical care, 42*(9), 851-859.
16. Torrance, G. W. (1976). Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-economic planning sciences, 10*(3), 129-136.
17. Farquhar, P. H. (1984). State of the art—utility assessment methods. *Management science, 30*(11), 1283-1300.
18. Kontodimopoulos, N., Argiriou, M., Theakos, N., & Niakas, D. (2011). The impact of disease severity on EQ-5D and SF-6D utility discrepancies in chronic heart failure. *The European Journal of Health Economics, 12*(4), 383-391.
19. Kularatna, S., Byrnes, J., Chan, Y. K., Carrington, M. J., Stewart, S., & Scuffham, P. A. (2017). Comparison of contemporaneous responses for EQ-5D-3L and Minnesota Living with Heart Failure; a case for disease specific multiattribute utility instrument in cardiovascular conditions. *International journal of cardiology, 227*, 172-176.
20. Wailoo, A. J., Hernandez-Alava, M., Manca, A., Mejia, A., Ray, J., Crawford, B., et al. (2017). Mapping to estimate health-state utility from non–preference-based outcome

measures: an ISPOR Good Practices for Outcomes Research Task Force Report. *Value in Health, 20*(1), 18-27.
21. Calxton, K., Martin, S., Soares, M., Rice, N., Spackman, E., Hinde, S., et al. (2013). Methods for the estimation of the NICE cost effectiveness threshold. Centre for Health Economics, University of York.
22. Drugs, C. A. f., & Health, T. i. (2006). Guidelines for economic evaluation of pharmaceuticals: Canada. *Ottawa: Canadian Agency for Drugs and Technologies in Health*.
23. Committee, P. B. A. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (PBAC). Version 5.0, September 2016. Canberra: Department of Health; 2016.
24. Kearns, B., Ara, R., Wailoo, A., Manca, A., Alava, M. H., Abrams, K., et al. (2013). Good practice guidelines for the use of statistical regression models in economic evaluations. *Pharmacoeconomics, 31*(8), 643-652.
25. Furber, G., Segal, L., Leach, M., & Cocks, J. (2014). Mapping scores from the Strengths and Difficulties Questionnaire (SDQ) to preference-based utility values. *Quality of Life Research, 23*(2), 403-411.
26. Ratcliffe, J., Flynn, T., Terlich, F., Stevens, K., Brazier, J., & Sawyer, M. (2012). Developing adolescent-specific health state values for economic evaluation. *Pharmacoeconomics, 30*(8), 713-727.
27. Gray, L. A., Alava, M. H., & Wailoo, A. J. (2017). Development of Methods for the Mapping of Utilities Using Mixture Models: Mapping the AQLQ-S to the EQ-5D-5L and the HUI3 in Patients with Asthma. *Value in Health, 21*(6), 748-757.
28. Varni, J. W., Burwinkle, T. M., & Lane, M. M. (2005). Health-related quality of life measurement in pediatric clinical practice: an appraisal and precept for future research and application. *Health and Quality of Life Outcomes, 3*(1), 34.
29. Edwards, B. (2014). Growing up in Australia: the longitudinal study of Australian children: entering adolescence and becoming a young adult. *Family Matters*(95), 5.
30. Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand journal of public health, 25*(5), 464-469.
31. Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research, 8*(1), 3-15.
32. Refaeilzadeh, P., Tang, L., & Liu, H. (2016). Cross-validation. *Encyclopedia of database systems*, 1-7.
33. Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry, 38*(5), 581-586.
34. Stevens, K. (2011). Assessing the performance of a new generic measure of health-related quality of life for children and refining it for use in health state valuation. *Applied health economics and health policy, 9*(3), 157-169.
35. Stevens, K. (2009). Developing a descriptive system for a new preference-based measure of health-related quality of life for children. *Quality of Life Research, 18*(8), 1105-1113.
36. StataCorp (2017). Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC.
37. Petrou, S., Rivero-Arias, O., Dakin, H., Longworth, L., Oppe, M., Froud, R., et al. (2015). The MAPS reporting statement for studies mapping onto generic preference-based outcome measures: explanation and elaboration. *Pharmacoeconomics, 33*(10), 993-1011.

38. Tosh, J. C., Longworth, L. J., & George, E. (2011). Utility values in National Institute for Health and Clinical Excellence (NICE) technology appraisals. *Value in Health, 14*(1), 102-109.
39. Round, J., & Hawton, A. (2017). Statistical alchemy: conceptual validity and mapping to generate health state utility values. *PharmacoEconomics-open, 1*(4), 233-239.
40. Schroeder, M. A., Lander, J., & Levine-Silverman, S. (1990). Diagnosing and dealing with multicollinearity. *Western Journal of Nursing Research, 12*(2), 175-187.
41. Dziuban, C. D., & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological bulletin, 81*(6), 358.
42. Tobias, S., & Carlson, J. E. (1969). Brief report: Bartlett's test of sphericity and chance findings in factor analysis. *Multivariate Behavioral Research, 4*(3), 375-377.
43. Brown, J. (2001). What is an eigenvalue? *JALT Testing & Evaluation SIG Newsletter, 5*(1).
44. Izquierdo, I., Olea, J., & Abad, F. J. (2014). Exploratory factor analysis in validation studies: Uses and recommendations. *Psicothema, 26*(3), 395-400.
45. Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*: MIT press.
46. Nelder, J. A., & Baker, R. J. (2004). Generalized linear models. *Encyclopedia of statistical sciences, 4*.
47. Masyn, K., Nathan, P., & Little, T. (2013). The Oxford Handbook of Quantitative Methods, Vol. 2: Statistical Analysis.
48. Manning, W. G., & Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of health economics, 20*(4), 461-494.
49. Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied statistics*, 15-14.
50. Pearson, E., & Please, N. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika, 62*(2), 223-241.
51. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398): John Wiley & Sons.
52. Glick, H. A., Doshi, J. A., Sonnad, S. S., & Polsky, D. (2014). *Economic evaluation in clinical trials*: OUP Oxford.
53. Basu, A. (2005). Extended generalized linear models: simultaneous estimation of flexible link and variance functions. *Stata Journal, 5*(4), 501.
54. Swearingen, C. J., Castro, M. M., & Bursac, Z. Inflated beta regression: Zero, one and everything in between. In *SAS global forum, 2012* (pp. 325-2012)
55. McDonald, J. F., & Moffitt, R. A. (1980). The uses of Tobit analysis. *The review of economics and statistics*, 318-321.
56. Longworth, L., Yang, Y., Young, T., Mulhern, B., Hernandez Alava, M., Mukuria, C., et al. (2014). Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. *Health Technology Assessment*.
57. Brennan, D. S., & Spencer, A. J. (2006). Mapping oral health related quality of life to generic health state values. *BMC health services research, 6*(1), 96.
58. Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics, 25*(3), 303-325.
59. Sullivan, P. W., & Ghushchyan, V. (2006). Mapping the EQ-5D index from the SF-12: US general population preferences in a nationally representative sample. *Medical Decision Making, 26*(4), 401-409.
60. McLachlan, G., & Peel, D. (2004). *Finite mixture models*: John Wiley & Sons.

61. Alava, M. H., Wailoo, A. J., & Ara, R. (2012). Tails from the peak district: adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. *Value in Health, 15*(3), 550-561.
62. Hernandez Alava, M., & Wailoo, A. (2015). Fitting adjusted limited dependent variable mixture models to EQ-5D. *Stata Journal, 15*(3), 737-750.
63. Grun, B., & Leisch, F. (2008). FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters.
64. Ratcliffe, J., Huynh, E., Chen, G., Stevens, K., Swait, J., Brazier, J., et al. (2016). Valuing the Child Health Utility 9D: using profile case best worst scaling methods to develop a new adolescent specific scoring algorithm. *Social Science & Medicine, 157*, 48-59.
65. Le, Q. A., & Doctor, J. N. (2011). Probabilistic mapping of descriptive health status responses onto health state utilities using Bayesian networks: an empirical analysis converting SF-12 into EQ-5D utility index in a national US sample. *Medical care*, 451-460.
66. Gray, A. M., Rivero-Arias, O., & Clarke, P. M. (2006). Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. *Medical Decision Making, 26*(1), 18-29.
67. Steyerberg, E. (2009). Validation of prediction models. In *Clinical Prediction Models* (pp. 299-311): Springer.
68. Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and psychological measurement, 34*(1), 111-117.
69. Collado-Mateo, D., Chen, G., Garcia-Gordillo, M. A., Iezzi, A., Adsuar, J. C., Olivares, P. R., et al. (2017). Fibromyalgia and quality of life: mapping the revised fibromyalgia impact questionnaire to the preference-based instruments. *Health and Quality of Life Outcomes, 15*(1), 114.
70. Teckle, P., McTaggart-Cowan, H., Van der Hoek, K., Chia, S., Melosky, B., Gelmon, K., et al. (2013). Mapping the FACT-G cancer-specific quality of life instrument to the EQ-5D and SF-6D. *Health and Quality of Life Outcomes, 11*(1), 203.
71. Kay, S., Tolley, K., Colayco, D., Khalaf, K., Anderson, P., & Globe, D. (2013). Mapping EQ-5D utility scores from the Incontinence Quality of Life Questionnaire among patients with neurogenic and idiopathic overactive bladder. *Value in Health, 16*(2), 394-402.
72. Jones, A. M., Lomas, J., Moore, P., & Rice, N. (2013). A quasi-Monte Carlo comparison of developments in parametric and semi-parametric regression methods for heavy tailed and non-normal data: with an application to healthcare costs. *Health Econometrics and Data Group Working Paper, 13*, 30.
73. Lamu, A. N., & Olsen, J. A. (2018). Testing alternative regression models to predict utilities: mapping the QLQ-C30 onto the EQ-5D-5L and the SF-6D. *Quality of Life Research, 27*(11), 2823-2839.
74. Rowen, D., Brazier, J., & Roberts, J. (2009). Mapping SF-36 onto the EQ-5D index: how reliable is the relationship? *Health and Quality of Life Outcomes, 7*(1), 27.
75. Goldsmith, K. A., Dyer, M. T., Buxton, M. J., & Sharples, L. D. (2010). Mapping of the EQ-5D index from clinical outcome measures and demographic variables in patients with coronary heart disease. *Health and Quality of Life Outcomes, 8*(1), 54.

Tables

Table 1: Respondent Characteristics

Variables	Sample statistics N = 6,898	Validation dataset N = 103
CHU9D health state utility value (observed)		
Mean (SD)	0.7976 (0.19)	0.6160 (0.22)
Range	-0.1059, 1	0.0560, 1
Utilities <0, n (%)	2 (0.03%)	0 (0.00%)
Utilities =1, n (%)	961 (13.93%)	1 (1.00%)
Utilities >0.9, n (%)	2,546 (36.90%)	14 (13.60%)
Utilities <0.5, n (%)	619 (8.97%)	36 (34.95%)
SDQ total score		
Mean (SD)	9.9400 (5.85)	20.3100 (5.14)
Range	0-32	9-32
SDQ subscales, Mean (SD)	Range: 1-10	0-10
Pro-social	7.8300 (1.72)	6.8000 (2.40)
Hyperactivity	3.7100 (2.32)	4.6400 (1.64)
Emotional symptoms	2.9100 (2.28)	5.3800 (2.61)
Peer problems	1.6800 (1.68)	5.8400 (1.60)
Conduct problems	1.6300 (1.58)	4.4500 (2.09)
Age (years)		
Mean (SD)	12.35 (2.05)	12.03 (3.14)
Range	10-15	5-17
Age categories, n (%)		
10-11 (B-Cohort)	3,567 (50.90%)	NA
14-15 (K-Cohort)	3,331 (49.09%)	NA
Female, n (%)	3,386 (49.08%)	54 (52.43%)
General health measure (self-reported), n (%)		
Excellent	3,271 (47.63%)	NA
Very good	2,604 (37.92%)	NA
Good	808 (11.77%)	NA
Fair	147 (2.14%)	NA
Poor	37 (0.54%)	NA

CHU9D: Child Health Utility-9D; NA: Not available; SD: Standard deviation; B-Cohort (or infant cohort) includes children born between March 2003 to February 2004; K-Cohort (or child cohort) includes children born between March 1999 to February 2000

Table 2: Predictive performance of models during ten-fold cross-validation

Models	Mean	Min	Max	MAE	MSE	MAE <0.5	MSE <0.5	MAE >0.9	MSE >0.9	Ranking[a]
Observed	0.7957	-0.1059	1							
Predictor set 1										
OLS	0.7975	0.4297	0.9963	0.1213	0.0249	0.3017	0.1057	0.1173	0.0212	30
GLM	0.7975	0.3808	0.9667	0.1210	0.0248	0.2976	0.1044	0.1173	0.0209	28
EEE	0.7975	0.3386	0.9742	0.1210	0.0248	0.2975	0.1046	0.1176	0.0209	29
ZOIB	0.7924	0.3914	0.9406	0.1222	0.0248	0.2983	0.1049	0.1249	0.0219	31
Tobit[b]	0.8124	0.4037	1	0.1199	0.0253	0.3020	0.1073	0.1025	0.0185	23
CLAD[b]	0.8200	0.4584	1	0.1198	0.0254	0.3266	0.1211	0.0993	0.0165	25
FMM	0.8047	0.5006	0.9677	0.1221	0.0253	0.3323	0.1233	0.1173	0.0194	32
MLOGIT	0.8169	0.3607	0.9347	0.1206	0.0256	0.3211	0.1211	0.1046	0.0163	26
Predictor set 2										
OLS	0.7975	0.4062	0.9939	0.1181	0.0238	0.2866	0.0972	0.1130	0.0202	21
GLM	0.7975	0.3557	0.9653	0.1178	0.0236	0.2825	0.0962	0.1132	0.0199	16
EEE	0.7974	0.2823	0.9702	0.1176	0.0236	0.2811	0.0960	0.1134	0.0198	15
ZOIB	0.7925	0.3701	0.9404	0.1192	0.0238	0.2852	0.0974	0.1210	0.0209	24
Tobit[b]	0.8122	0.3817	1	0.1169	0.0241	0.2863	0.0985	0.0988	0.0177	12
CLAD[b]	0.8158	0.4146	1	0.1169	0.0241	0.3063	0.1090	0.0971	0.0163	14
FMM	0.8049	0.4815	0.9675	0.1191	0.0242	0.3202	0.1156	0.1135	0.0185	27
MLOGIT	0.8168	0.3412	0.9375	0.1176	0.0243	0.3061	0.1122	0.1009	0.0155	19
Predictor set 3										
OLS	0.7977	0.3831	0.9962	0.1163	0.0232	0.2758	0.0919	0.1109	0.0198	7
GLM[c]	**0.7977**	**0.3204**	**0.9598**	**0.1156**	**0.0230**	**0.2689**	**0.0900**	**0.1100**	**0.0191**	**1**
EEE	0.7982	0.2473	0.9704	0.1155	0.0230	0.2983	0.1049	0.1249	0.0219	18
ZOIB	0.7934	0.3591	0.9507	0.1171	0.0231	0.2768	0.0934	0.1177	0.0201	11
Tobit[b]	0.8119	0.3750	1	0.1154	0.0236	0.2766	0.0936	0.0973	0.0174	2
CLAD[b]	0.8142	0.3583	1	0.1161	0.0240	0.2797	0.0967	0.0983	0.0173	8
FMM	0.8055	0.4961	0.9762	0.1175	0.0237	0.3156	0.1127	0.1107	0.0180	20
MLOGIT	0.8169	0.3099	0.9538	0.1151	0.0234	0.2991	0.1066	0.0969	0.0149	4
Predictor set 4										
OLS	0.7976	0.3870	0.9849	0.1166	0.0232	0.2763	0.0920	0.1115	0.0198	9
GLM	0.7976	0.3420	0.9615	0.1163	0.0231	0.2733	0.0915	0.1116	0.0194	5
EEE	0.7976	0.2780	0.9661	0.1162	0.0231	0.2852	0.0974	0.1210	0.0209	17
ZOIB	0.7931	0.3489	0.9402	0.1176	0.0232	0.2757	0.0929	0.1189	0.0203	13
Tobit[b]	0.8121	0.3657	1	0.1159	0.0236	0.2768	0.0938	0.0975	0.0173	3
CLAD[b]	0.8141	0.3813	1	0.1160	0.0237	0.2809	0.0963	0.0971	0.0169	6
FMM	0.8053	0.4719	0.9611	0.1178	0.0237	0.3135	0.1115	0.1121	0.0181	22
MLOGIT	0.8169	0.3239	0.9399	0.1162	0.0237	0.2984	0.1079	0.0992	0.0151	10

Acronyms: CLAD: Censored Least Absolute Deviation; EEE: Extended Estimating Equations; FMM: Finite Mixture Model; GLM: Generalised Linear Model; MAE: Mean Absolute Error; MLOGIT: Multinomial Logistic Regression; MSE: Mean Squared Error; OLS: Ordinary Least Squares; ZOIB: Zero-One-Inflated Beta binomial

Explanatory variables for: Predictor set 1: Total SDQs and gender; Predictor set 2: SDQs subscales (social, hyperactivity, emotional, peer, conductive) and gender; Predictor set 3: Individual SDQ items (categorical) and gender; and Predictor set 4: Individual SDQ items (continuous) and gender.

The best-ranked model is highlighted in bold

[a]Some ranks were rearranged (based on lowest possible overall MSE values) when their overall rankings were equal

[b]Results were truncated to 1 whenever the predicted utilities exceeded 1.

[c]Summary statistics of GLM 3 using full sample- Mean (SD): 0.7976 (0.12); Min :0.2000; Max: 0.9600.

Total predicted latent utilities exceeding one for TOBIT models during cross-validation: Tobit 1: 142; Tobit2: 164; Tobit 3: 157; Tobit 4: 257. Total predicted latent utilities exceeding one for CLAD models during cross-validation: CLAD 1: 51; CLAD 2: 56; CLAD 3: 73; CLAD 4: 50.

Table 3: Coefficients and standard errors from the best performing model

Predictor variables	Coefficients	Robust Standard errors
SDQ1_2	-0.0498	0.0299
SDQ1_3	-0.0563	0.0298
SDQ2_2	0.0102	0.0051
SDQ2_3	-0.0011	0.0072
SDQ3_2	0.0381	0.0047
SDQ3_3	0.0671	0.0080
SDQ4_2	0.0055	0.0077
SDQ4_3	0.0020	0.0078
SDQ5_2	0.0064	0.0048
SDQ5_3	0.0127	0.0083
SDQ6_2	0.0280	0.0053
SDQ6_3	0.0192	0.0098
SDQ7_2	0.0057	0.0124
SDQ7_3	-0.0064	0.0127
SDQ8_2	0.0602	0.0048
SDQ8_3	0.0964	0.0070
SDQ9_2	-0.0154	0.0170
SDQ9_3	-0.0158	0.0171
SDQ10_2	0.0068	0.0050
SDQ10_3	0.0037	0.0079
SDQ11_2	-0.0193	0.0193
SDQ11_3	-0.0337	0.0181
SDQ12_2	0.0023	0.0090
SDQ12_3	-0.0191	0.0232
SDQ13_2	0.0649	0.0059
SDQ13_3	0.1137	0.0108
SDQ14_2	-0.0207	0.0111
SDQ14_3	-0.0365	0.0114
SDQ15_2	0.0233	0.0050
SDQ15_3	0.0287	0.0074
SDQ16_2	0.0273	0.0048
SDQ16_3	0.0353	0.0068
SDQ17_2	-0.0103	0.0169
SDQ17_3	-0.0125	0.0164
SDQ18_2	0.0230	0.0054
SDQ18_3	0.0157	0.0088
SDQ19_2	0.0257	0.0064
SDQ19_3	0.0138	0.0104
SDQ20_2	0.0052	0.0069
SDQ20_3	-0.0074	0.0076
SDQ21_2	0.0094	0.0110
SDQ21_3	-0.0115	0.0116
SDQ22_2	0.0129	0.0091
SDQ22_3	0.0009	0.0238
SDQ23_2	0.0139	0.0045
SDQ23_3	0.0070	0.0090
SDQ24_2	0.0266	0.0048
SDQ24_3	0.0451	0.0081
SDQ25_2	-0.0251	0.0078
SDQ25_3	-0.0510	0.0088
Female	0.0249	0.0044
Constant	0.4291	0.0444

SDQ= Strength and Difficulties Questionnaire.

The Best model was GLM 3 that included categorical SDQ items and gender (Female==1) as predictors. Disutility (=1-Utility) was used as the dependent variable. Poisson family and a power (1/2) link function were selected based on tests.